Genetic drift model as rational inference

Ryota Morimoto

Macroscopic worldview

Genetic drift is considered one of the major evolutionary factors. It refers to chance fluctuation of gene frequency. To express genetic drift mathematically, probability is an indispensable concept. This raises a philosophical question: What is the appropriate interpretation of probability in drift model?

First of all, we will make sure of macroscopic worldview. Imagine that a woman and a man are walking over the bridge without handrail between mountains. The woman follows behind the man, and if the woman pushes the man' back, what happens? Logically there are infinite possibilities, e.g., flying away, moving backwards and so on. But in our actual world there is only one trajectory, i.e., falling down. And Newtonian mechanics describes it.

Pierre Laplace (1814) formulates Newtonian worldview. Laplace assumes an intelligence, which is called Laplace's demon after his name. Laplace supposes demon has complete information and perfect calculating power. Demon knows initial state of the system, and substituting it into the laws of Newtonian mechanics, then it can calculate an unique final state with perfect computation. For an intelligence, as Laplace says, "nothing would be uncertain and the future, as the past, would be present to its eyes" (ibid., p.4). So the world is deterministic, and there is no chance-like or probabilistic event in the world. In deterministic worldview, probability concept is interpreted as our ignorance, not represented the world. Laplace says, "probability is relative, in part to this our ignorance" (ibid., p.6). This is deterministic worldview in material world.

Let us turn our attention to the biological world. There seems to be chance-like event, like natural selection, random drift, etc. I will focus only on random genetic drift in this paper. Random drift is often explained as an example of random sampling. Assume that diploid organisms in some population have either allele A or a on some locus and that the frequency of A is p in the parental generation. There are a large number of gametes at the time of reproduction of parental generation, but only $2N$ gametes are sampled from them in offspring generation. So there will be N individuals in offspring generation. The probability $p(i)$ that the number of allele A is i in the next generation is expressed by

$$p(i) = \binom{2N}{i} p^i (1 - p)^{2N - i} . \tag{1}$$

This is a standard genetic drift model called Wright-Fisher model. Notice that probability concepts appear in this equation.

What is the appropriate interpretation of the probability concept in drift model? Alex Rosenberg (1994) provides an answer to this question. I will summarize his argument briefly. Rosenberg says evolution including drift occurs at the individual-level. Individual organism behaves at the macroscopic level. From Newtonian mechanical point of view, macroscopic phenomena are deterministic, as Laplace formulates. Then evolutionary phenomena are deterministic. According to deterministic worldview, as I explained, probability concepts in drift model is interpreted as our ignorance. This is Rosenberg's answer to the problem of interpretation of probability concepts in drift model.

Population-level phenomena

There are some critiques of Rosenberg's argument. Let me introduce a critique of first sentence of his argument, i.e., evolution including drift occurs at the individual level. Walsh, Lewens, and Ariew (2002) criticize Rosenberg's argument. They raise simple questions. When you toss a coin ten times, 6 head and 4 tails is more likely than 9 head and 1 tail. And when you toss the coin, 9 heads and 1 tail is more likely than 99 heads and 1 tail. In these cases, similar probability distributions are obtained whoever tries, so these are objective phenomena to be explained. Why such phenomena happens? Walsh et al. claim dynamical or Newtonian mechanical account at individual level cannot explain such phenomena, but statistical account can. So there exist population-level phenomena. Citing some examples of drift case, like Hagedoorn effect, Wright effect and bottle neck effect, Walsh et al. claim drift can occur at not individual but population level. I agree with them. Remember that Francis Galton created the quincunx device to demonstrate bell-shaped curve of the normal distribution. Whoever drops the balls, the similar probability distribution can be obtained. It seems too difficult for Rosenberg to explain such population-level phenomena on the base of Newtonian mechanics.

Let us turn our attention to Wright-Fisher drift model. It says the probability that the number of allele A is i in the next generation is expressed binominal distribution in equation (1). In the standard derivation of drift model we need some assumptions, for example, random gamete sampling, sexual reproduction, constant population size, no selection. Among these assumptions I focus on the assumption of random gamete sampling. This assumption means each gamete has an ‘equal probability’ of sampling. Are all gametes the same? Is this equality assumption empirically grounded? At the molecular level empirical data shows there are selectively neutral or nearly neutral alleles (Kimura 1983). But how about organism or population level? No consensus exists yet. Application of Wright-Fisher model is not restricted to the molecular level. It can also be applied to higher-level phenomena.

Fortunately, we can derive Wright-Fisher drift model without equality or neutrality assumption. Morimoto (2009a) adopts Jayens’ works to drift model. Statistical physicist Edward Jaynes (1957a; 1957b) derives equal probability by the use of the maximum entropy principle, which is a method of information theory. According to this principle, when entropy is maximum, we can make rational inference. Jaynes adopts maximum entropy principle to statistical mechanics. Standard approaches to statistical mechanics are based on the postulate of equal a priori probability, which is introduced by statistical physicist Richard Tolman (1938). It says that for an isolated system in equilibrium, it is found with equal porbability in any of its accessible microstates. On the other hand, Jaynes shows the standard formalisms in statistical mechanics can be derived without assuming ‘equal probability’ by using the maximum entropy principle and claimed that statistical mechanics is a consequence of rational inference.

I will summarize Jaynes’ work by applying it to gamete sampling case. What we want to know is the probability that i gametes are sampled from $2N$ gametes. Now we know the sum of all probabilities equals to 1. This is one of axioms of probability theory. Suppose that this is all information we have. Notice that we have no information about equality or neutrality. And then we maximize information entropy which is derived by Claude Shannon (1948). Subject to partial information D_1 (in this case axiom of probability) we maximize entropy, that is, utilize our information most efficiently, then we obtain equal sampling probability

$$p(i \mid D_1) = \frac{i}{2N} \quad . \tag{2}$$

This equation expresses that sampling probability of each gametes is 'equal' (see Appendix 1 for detailed derivation). In this derivation equal probability is not an assumption, but a consequence of rational inference. Then Jaynes says there is no need for the principle of indifference nor of a priori probability in statistical mechanics.

Jaynes' work is not about biology but about physics. Morimoto (2009a) adopts his work to drift model. There are many gametes in one generation and in the next generation only finite $2N$ gametes are drawn, so there are N individuals because of assumption of diploid organism. Suppose that in parental generation the frequency of gamete A is p and that in offspring generation the number of A is i. Let us number the gamete and define random variable x_k as follows. x_k is 1 if the number k allele is A, and x_k is 0 if the number k allele is a. Let p_k be the probability that k allele is A, and we don't know what this probability is. In offspring generation, the number of allele A is i and the expected number of allele A in offspring generation is

$$\textstyle\sum_{k=1}^{2N} x_k p_k = 2Np \ . \tag{3}$$

and call this information D_2.

In this situation we know axiom of probability which is denoted D_1 above and we have further information, that is, expectation of the number of gametes A in offspring generation, which is denoted D_2. Here again we don't assume equal probability. Subject to these partial information (D_1 & D_2), we maximize information entropy function by the use of maximum entropy principle, we obtain

$$p(i \mid D_1 \;\&\; D_2) = \binom{2N}{i} p^i (1 - p)^{2N - i} \tag{4}$$

This equation is identical to Wright-Fisher model in equation (1) (see Appendix 2 for detailed derivation).

By the use of maximum entropy principle we can derive drift model without equality assumption. Therefore complete or full information which Laplace's demon could have, is not needed to derive drift model. However this doesn't mean that drift model is incomplete or that it can't capture the real aspect of biological world. For we use observable and objective information about population, e.g., frequency of *A*, population size and so on. Such information reflects some aspects of reality.

Let us summarize the derivation of drift model by the use of maximum entropy principle. Maximum entropy principle is a tool for rational inference from partial information. Here we have two kinds of information, i.e., axiom of probability and expectation. And notice that we don't assume equal probability and we use information about objective properties of population. Under these constraints we can derive drift model by using maximum entropy principle.

Bayesian interpretation

To clarify the meaning of probability concepts in drift model, I will explore maximum entropy principle further. Economist and statistician Arnold Zellner (1988) derives Bayes' theorem by using maximum entropy principle. He says updating information by using MEP is an optimal information processing.

Suppose that there are some hypothesis H_k and data D and we have information in hypothesis $p(H_k)$, information in data $p(D)$, and information in data given hypothesis $p(D \mid H_k)$. And by using some information processing rule, we get output information, that is, information in hypothesis given data $p(H_k \mid D)$. Zellner claims that when we use different information processing rule like rule of adding irrelevant information or rule of decreasing information, we get different output information. And to minimize difference between input and output information, that is, to minimize information loss, is an an optimal information processing rule. Here again we know the sum of all probabilities equals to 1 (D_1). If we maximize entropy, that is, we minimize information loss, subject to partial information D_1, we obtain

$$p(H_k \mid D_1) = \frac{p(D_1 \,\&\, H_k)\, p(H_k)}{\sum_k p(D_1 \,\&\, H_k)\, p(H_k)} \quad . \tag{5}$$

This equation is identical to Bayes' theorem (see Appendix 3 for detailed derivation). To maximize information entropy means to utilize our information most efficiently, that is, to minimize information loss. Therefore Bayes' theorem can be derived from maximum entropy principle.

From information theoretical point of view, Bayes' theorem is a result of optimal information processing rules. When we update partial information optimally, we can derive drift model. If we just know one of axioms of probability, we can obtain equal probability by using MEP. If we have additional information, we can derive drift model by updating information optimally.

I have attempted to interpret the probability concepts in genetic drift model from information theoretical point of view. If my attempt meets with success, drift model can be inferred by updating partial information optimally. To derive drift model, we don't need complete of full information but partial one. We just know axiom of probability and the expectation. To derive drift model other assumptions including equality one is not needed. Even if we could have further information, we dare to dismiss it.

Moreover I show that probability concepts in drift model can be interpreted as Bayesian. Consequences of inference depend on what we know. Bayesian interpretation is one of the subjective interpretation. Namely, probability may change depending on what we know. If we just know axiom of probability, then we obtain equal probability by the use of maximum entropy principle. And if we have additional information of expectation, we get genetic drift model by using it. However, even if we put this interpretation on drift model, it doesn't mean drift model fails to capture objective features. In fact, as we saw, in derivation of drift model we use objective properties of population, like population size, frequency of allele A. So probability concepts in drift model can be interpreted as not just our ignorance as Rosenberg says, but rational inference from partial information.

References

Bouchard, F. and Rosenberg, A. 2004. Fitness, Probability and the Principles of Natural Selection. *British Journal for the Philosophy of Science* 55: 693-712.

Graves, L., Horan, B., and Rosenberg, A. 1999. Is Indeterminism the Source of the Statistical Character of Evolutionary Theory. *Philosophy of Science* 66: 140-157.

Jaynes, E. T. 1957a. Information Theory and Statistical Mechanics. *Physical Review*, 106: 620-630.

Jaynes, E. T. 1957b. Information Theory and Statistical Mechanics II. *Physical Review*, 108: 171-190.

Jaynes, E T. 1968. Prior Probability. IEE Transactions On Systems. *Science and Cybernetics* 4: 227-241.

Jaynes, E. T. 1988. The Relation of Bayesian and Maximum Entropy Methods. *Maximum-Entropy and Bayesian Methods in Science and Engineering*: 1: 25-29.

Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. New York: Cambridge University Press.

Laplace, P. 1814 [1951]. *A Philosophical Essay on Probabilities*. London, Dover.

Millstein, R. 2006. Natural Selection as a Population-Level Causal Process. *British Journal for the Philosophy of Science* 57: 627-653.

Morimoto, R. 2008. Information Theory and Natural Selection. *Annals of Japan Association for Philosophy of Science* 16: 57-73.

Morimoto, R. 2009a. Genetic Drift and Information Theory. *Journal of Biological Science, Japan* 60: 197-204.

Morimoto, R. 2009b. Interpretation of Probability in Evolutionary theory -Possibility of Bayesian Interpretation. *Journal of the Philosophy of Science Society, Japan* 42:83-96.

Rosenberg, A. 1994. *Instrumental Biology, or The Disunity of Science*. Chicago, IL: The University of Chicago Press.

Shannon, C. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379-423.

Sober, E. 1984. *The Nature of Selection.* Cambridge, MA: The MIT Press.

Tolman, R. 1938. *The Principles of Statistical Mechanics.* Oxford University Press (reprinted Dover Publication 1979).

Walsh, D., Lewens, T., and Ariew, A. 2002. The Trials of Life: Natural Selection and Random Drift. *Philosophy of Science* 69: 452-473.

Zellner, A. 1988. Optimal Information Processing and Bayes's Theorem. *The American Statistician* 42: 278-280.

Appendix 1. Equal probability

Let p_k (k = 1,2,…2N) stand for probability that gamete k is sampled from 2N ones. Suppose that we are not given the value of probability p_k , but we just know one of the axioms of probability

$$\textstyle\sum_{k=1}^{2N} p_k = 1 \ . \tag{A 1.1}$$

We call this information data 1; D_1. In addition, Shannon (1948) proved that the quantity, which is positive, which increases with increasing uncertainty, and which is additive for independent source of uncertainty, is the information entropy function

$$H = -\textstyle\sum_{k=1}^{2N} p_k \log p_k \ . \tag{A 1.2}$$

In deriving p_k on the basis of partial information, we ought to use the probability which has maximum entropy subject to whatever we know. We show that maximization of H leads to probability p_k by equating the derivation to 0 subject to the constraint. Maximizing H yields

$$dH = 0 \ . \tag{A 1.3}$$

Now there is a constraint; one of the axioms of probability (A 1.1). Differentiating this gives

$$\textstyle\sum_{k=1}^{2N} dp_k = 0 \ . \tag{A 1.4}$$

We maximize H by using the method of maximum entropy principle. We obtain

$$dH - \alpha \textstyle\sum_{k=1}^{2N} dp_k = 0 \, , \tag{A 1.5}$$

where α is a Lagrange multiplier. Now differentiating (A 1.2) gives

$$\frac{dH}{dp_k} = -\sum_{k=1}^{2N} (1 + \log p_k) . \tag{A 1.6}$$

Substituting this into (A 1.5) yields

$$-\sum_{k=1}^{2N} \left[(1 + \log p_k) + \alpha \right] dp_k = 0 . \tag{A 1.7}$$

All these coefficients of dp_k must be 0 in order to satisfy this identical equation. Then

$$(1 + \log p_k) + \alpha = 0 . \tag{A 1.8}$$

Transforming this equation yields

$$p_k = \exp(- \alpha - 1) , \tag{A 1.9}$$

Substituting this into (A 1.1) becomes

$$\sum_{k=1}^{2N} p_k = \sum_{k=1}^{2N} \exp(- \alpha - 1) = \exp(- \alpha - 1) \sum_{k=1}^{2N} = 2N \exp(- \alpha - 1) = 1 . \tag{A 1.10}$$

Then

$$\exp(- \alpha - 1) = \frac{1}{2N} . \tag{A 1.11}$$

Substituting this into (A 1.9) gives

$$p_k = \frac{1}{2N} . \tag{A 1.12}$$

Thus probability p_k can be derived by using maximum entropy principle. Further, when i gametes are sampled, we obtain

$$p(i \mid D_i) = \frac{i}{2N} \ . \qquad \text{(A 1.13)}$$

Appendix 2. Wright-Fisher drift model

We will derive Wright-Fisher drift model by using maximum entropy principle. Initially, assume that diploid organisms in some population have either allele *A* or *a* on a specific locus and that the frequency of *A* is *p* in the parental generation. There are a large number of gametes at the time of reproduction of parental generation, but we suppose that only 2*N* gametes are sampled from them in offspring generation. So there will be *N* individuals in offspring generation. Now we want to know the probability that the number of allele *A* is *i* in the next generation. Let p_k (k = 1,2,…2N) stand for the probability that *k* allele in the offspring generation is *A* and we don't know what it is. Again suppose that we know normalization. As we see in A1 above, this is one of the axioms of probability theory;

$$\textstyle\sum_{k=1}^{2N} p_k = 1 \ , \qquad \text{(A 2.1)}$$

and call this information D_1. Further, we know expected number of allele *A* in offspring generation.

Let us number each allele in this time from 1 to 2*N* and define random variables x_k as follows. x_k is 1 if the number *k* allele is *A*, and x_k is 0 if the number *k* allele is *a*. In offspring generation, the number of allele *A* is *i*, then

$$\textstyle\sum_{k=1}^{2N} x_k = i \ . \qquad \text{(A 2.2)}$$

And expected number of allele *A* in offspring generation is

$$\textstyle\sum_{k=1}^{2N} x_k p_k = 2Np \ . \qquad \text{(A 2.3)}$$

and call this information D_2. Suppose that we know only D_1 and D_2, then here are two constraints; normalization and expectation. Differentiating each constraint gives

$$\sum_{k=1}^{2N} dp_k = 0\,, \tag{A 2.4}$$

and

$$\sum_{k=1}^{2N} x_k\, dp_k = 0\,. \tag{A 2.5}$$

By maximizing H under these constraints, we obtain

$$dH - \alpha\sum_{k=1}^{2N} dp_k - \beta\sum_{k=1}^{2N} x_k\, dp_k = 0\,, \tag{A 2.6}$$

where α and β are Lagrange multipliers. Substituting entropy H into (A2.6) yields

$$-\sum_{k=1}^{2N}\left[(1+\log p_k)+\alpha+\beta x_k\right]dp_k = 0\,. \tag{A 2.7}$$

and then

$$1+\log p_k + \alpha + \beta x_k = 0\,. \tag{A 2.8}$$

Transforming this equation yields

$$p_k = \exp(-\alpha - \beta x_k - 1). \tag{A 2.9}$$

Substituting this equation into (A2.1) yields

$$\sum_{k=1}^{2N} p_k = \sum_{k=1}^{2N}\exp(-\alpha-\beta x_k - 1) = \exp(-\alpha-1)\sum_{k=1}^{2N}\exp(-\beta x_k) = 1\,. \tag{A 2.10}$$

Then we obtain

$$\exp(-\alpha-1) = \frac{1}{\sum_{k=1}^{2N}\exp(-\beta x_k)} \tag{A 2.11}$$

Substituting this into (A2.9) becomes

$$p_k = \frac{\exp(-\beta x_k)}{\sum_{k=1}^{2N} \exp(-\beta x_k)} = \frac{\exp(-\beta x_k)}{[\exp(-\beta\ ) + 1]^{2N}} \tag{A 2.12}$$

Next, to erase multiplier *β*, we ought to differentiate entropy function *H* subject to the two constraints. Entropy function and normalization are not the function of multiplier *β*. So we only differentiate expected value with respect to *β*, then we obtain

$$\sum_{k=1}^{2N} x_k\, p_k - 2Np = 0\ . \tag{A 2.13}$$

Substituting (A2.12) into this equation yields

$$\frac{\sum_{k=1}^{2N} x_k\ \exp(-\beta x_k)}{\sum_{k=1}^{2N} \exp(-\beta x_k)} = 2Np \tag{A 2.14}$$

The left-hand side of this equation is transformed to

$$\frac{\sum_{k=1}^{2N} x_k\ \exp(-\beta x_k)}{\sum_{k=1}^{2N} \exp(-\beta x_k)} = \frac{d}{d\beta} \log\left[\sum_{k=1}^{2N} \exp(-\beta x_k)\right]$$

$$= \frac{d}{d\beta} \log\left(\exp(-\beta x_k) + 1\right)^{2N} = \frac{2N}{\exp(\beta) + 1}\ . \tag{A 2.15}$$

Then multiplier *β* is

$$\beta = \log\frac{1-p}{p}\ . \tag{A 2.16}$$

Substituting this into (A2.12) yields

$$p_k = \frac{\left(\dfrac{p}{1-p}\right)^{x_k}}{\left(\dfrac{p}{1-p} + 1\right)^{2N}} = p^i(\ 1-p\ )^{2N-i}\ . \tag{A 2.17}$$

This is the probability that k allele in the offspring generation is A. Summing (A2.17) from $k = 1$ to $2N$ and substituting (A2.2) into the result, we obtain

$$p(i \mid D_1 \ \& \ D_2) = \binom{2N}{i} p^i (1 - p)^{2N - i} \qquad \text{(A 2.18)}$$

This result is identical to Wright-Fisher drift model. Therefore it can be derived by using the method of maximum entropy principle.

Appendix 3. Bayes' theorem

Let H_k (k = 1,2, …, n) stand for hypothesis and D for data. There is a need to measure information in the input and output probability functions. The following measures will be employed;

Information in $$p(H_k) = -\sum_{k=1}^{n} p(H_k \mid D) \log p(H_k) \qquad \text{(A3.1)}$$

Information in $$p(D) = -\sum_{k=1}^{n} p(H_k \mid D) \log p(D) = -\log p(D) \qquad \text{(A3.2)}$$

Information in $$p(D \mid H_k) = -\sum_{k=1}^{n} p(H_k \mid D) \log p(D \mid H_k) \qquad \text{(A3.3)}$$

Information in $$p(H_k \mid D) = -\sum_{k=1}^{n} p(H_k \mid D) \log p(H_k \mid D) \qquad \text{(A3.4)}$$

In each case, information is given as an average of a log probability function with p(H_k |D) used as a weight function. The difference between the output and input information is represented as

$$L = \sum_i p(H_k \mid D) \log p(H_k \mid D) + \log p(D) - \sum_k p(H_k \mid D) \log p(D \mid H_k)$$
$$- \sum_k p(H_k \mid D) \log p(H_k) \, . \qquad \text{(A3.5)}$$

According to an optimal information processing rule, the output information should be as close as possible to the input information and ideally equal to it. To minimize information loss with this rule, we ought to minimize (A3.5). Here again, suppose that we know only that the sum of all probabilities equals to 1;

$$\textstyle\sum_k p\,(H_k \mid D\,) = 1\ . \tag{A3.6}$$

Minimize (A3.5) subject to the condition (A3.6) by using the method of maximum entropy principle, we obtain

$$\frac{d\left[L + \alpha\left\{\sum_k dp\,(H_k \mid D\,) - 1\right\}\right]}{dH_k} = 0\ , \tag{A3.7}$$

where *α* is a Lagrange multiplier. The left-hand side of this equation is transformed to

$$\begin{aligned}\frac{d}{dH_k}\Big[&\textstyle\sum_k p\,(H_k \mid D\,)\log p\,(H_k \mid D\,) + \log p\,(D) - \sum_k p\,(H_k \mid D\,)\log p\,(D \mid H_k\,)\\ &\textstyle - \sum_k p\,(H_k \mid D\,)\log p\,(H_k\,) + \alpha\left\{\sum_k dp\,(H_k \mid D\,) - 1\right\}\Big]\\ = \ &\textstyle\sum_k\left[\log p\,(H_k \mid D\,) - \log p\,(D \mid H_k\,) - \log p\,(H_k\,) + \alpha\right] + 1 - 1\\ = \ &\textstyle\sum_k\left[\log p\,(H_k \mid D\,) - \log p\,(D \mid H_k\,) - \log p\,(H_k\,) + \alpha\right]\ .\end{aligned}$$

Then we get

$$\log p\,(H_k \mid D\,) = \log p\,(D \mid H_k\,) + \log p\,(H_k\,) - \alpha\ . \tag{A3.8}$$

Substituting this equation into (A3.6) and differentiating respect to H_k yields

$$\textstyle\sum_k \log p\,(H_k \mid D\,) = \sum_k \log p\,(D \mid H_k\,) + \sum_k \log p\,(H_k\,) - \alpha = 0\ . \tag{A3.9}$$

So we obtain

$$\alpha = \textstyle\sum_k \log p\,(D \mid H_k\,)\log p\,(H_k\,) \tag{A3.10}$$

Substituting this into (A3.8) gives

$$p(H_k \mid D) = \frac{p(D \mid H_k) \log p(H_k)}{\sum_k (D \mid H_k) \log p(H_k)} \tag{A3.11}$$

This equation is identical to Bayes' theorem. Therefore we can derive Bayes' theorem from maximum entropy principle.